

# Scheduling Reserved Traffic in Input-Queued Switches: New Delay Bounds via Probabilistic Techniques

Matthew Andrews and Milan Vojnović\*

EPFL I&C Technical Report IC/2002/62, September 2002

## Abstract

We consider the problem of providing delay bounds to reserved traffic in high-speed input-queued switches. We assume that the matrix of bandwidth demands is known and we use the now standard approach of decomposing this matrix into a convex combination of permutation matrices. Our problem therefore reduces to the problem of constructing a schedule for these permutation matrices.

In this paper we derive delay bounds for four algorithms that are based on probabilistic techniques. For each algorithm we first place tokens randomly in continuous time for each permutation matrix. If the  $n$ th token that appears corresponds to permutation matrix  $M_k$  then we schedule matrix  $M_k$  in the  $n$ th time slot. The algorithms differ in how the random token processes are defined. For two of the algorithms we are able to perform a derandomization so as to obtain deterministic schedules.

We show through numerical computation that in many situations the resulting delay bounds are smaller than the previously best-known delay bounds of Chang, Chen, and Huang [1].

**Index terms:** Input-queued switches. Decomposition-based schedules. Delay bounds.

## 1 Introduction

In recent years there has been a great deal of work on scheduling algorithms for input-queued switches. The key feature of an input-queued switch is that at each time step, each input can be connected to at most one output and each output can be connected to at most one input. The aim of the scheduler is

---

\*Author Affiliations: Matthew Andrews, Bell Laboratories, Lucent Technologies, Murray Hill, NJ <andrews@research.bell-labs.com>, Milan Vojnović, EPFL, CH-1015 Lausanne, Switzerland <mi-lan.vojnovic@epfl.ch>. Work performed in part while the second author was visiting Bell Labs.

to determine how to configure the switch at each time step so as to provide high throughput and low delays for the arriving packets.

Most of the previous work has concentrated on providing *stability* or *100% throughput* for the switch. A scheduler is said to be stable if the queues remain bounded as long as the load on each port is less than the capacity of that port. Algorithms that provide stability generally fall into two categories depending on whether or not we know the arrival rates for each input-output pair in advance. For the case in which we do know the arrival rates we can decompose the rate matrix into a convex combination of permutation matrices. If the scheduler configures the switch according to this decomposition then we have stability. We refer to these schedulers as *decomposition-based* schedulers (e.g. [1, 2, 3]). Specific algorithms for performing the decomposition can be derived from results of Birkhoff [4] and von Neumann [5]. Note that the schedule can be computed in advance of the arriving traffic and hence it is acceptable for the calculation to have significant complexity.

For the case of high-speed optical switches it is reasonable to assume that the arrival rates are known to us because such switches are likely to be deployed in the core of networks where traffic engineering using MPLS is becoming more prevalent. Switches that support MPLS must be able to provide *bandwidth guarantees* for certain input-output pairs. For each MPLS path that passes through input  $i$  and output  $j$  on a switch, the switch will be required to *reserve* bandwidth for the path between these two ports. Another justification that the input rates are known can be found with Expedited Forwarding (EF) in the context of differentiated services [6]. There it is commonplace to assume the network is engineered such that the load of EF traffic at each node is bounded by some configured value.

For the case in which we do not know the arrival rate matrix in advance then most of the stable schedulers set up a bipartite graph whose edge weights correspond to the queue sizes of the corresponding input queues. At each time step the scheduler finds an (approximate) maximum-weight matching in this graph and configures the switch accordingly. These schedulers are sometimes known as maximum weight matching (MWM) type schedulers (e.g. [7, 8, 9, 10]). Note that an MWM scheduler must operate in real-time since it needs access to queue information and hence it must have extremely low complexity.

More recently, attention has been paid to the problem of minimizing the *delay* experienced by packets passing through an input-queued switch. Leonardi et al. [11] analyzed the MWM algorithm and showed that the mean delay through an arbitrary pair of input-output ports of a switch with  $I$  input and  $I$  output ports, uniformly loaded to  $\rho < 1$ , is bounded by  $(I - \rho)/(1 - \rho)$ . A related work is that of Shah and Kopikare [12] who observe that for uniform Bernoulli arrivals to the switch with the scheduling policy that at each time slot takes a matching uniformly at random from the entire set of  $I!$  matchings, the expected delay is  $(I - 1)/(1 - \rho)$ . Note that this is smaller than the bound obtained in [11].

In [1] Chang, Chen, and Huang showed how to derive worst-case deterministic delay bounds for a Birkhoff-von Neumann schedule in which the permutation matrices are scheduled according to Packetized Generalized Processor Sharing (PGPS) [13] such that at the time a matrix is served it is placed as a new arrival into the PGPS system.

In this paper we consider delay bounds for decomposition-based schedules. We show that by using probabilistic techniques we are able to tighten the bounds of [1] for the worst case input-output pairs in many scenarios. We use the term “probabilistic techniques” rather than “randomized algorithms” since for two of our algorithms we are able to derandomize the random processes so as to obtain *deterministic algorithms*.

It is interesting to observe that for a node to support EF it needs to conform to a rigorous definition of the per-hop-behavior (Packet Scale Rate Guarantee with rate  $r$  and latency  $e$ ; see [14]). Our work can be viewed as calculating what the value of the latency  $e$  would be for an input-queued switch. Before we present our algorithms and results in detail we must present some notation and our goals.

## Assumptions and Notation

We consider an  $I \times I$  switch. Let  $\rho_{ij}$  be the bandwidth that needs to be reserved between input  $i$  and output  $j$ , normalized by the link rate. Let  $M$  be the matrix whose  $ij$  entry is  $\rho_{ij}$ . We refer to  $M$  as the *rate matrix*. For the majority of this paper we consider the case in which  $M$  is a *doubly stochastic* matrix, i.e.  $\sum_i \rho_{ij} = 1$  and  $\sum_j \rho_{ij} = 1$ . This corresponds to the case in which the entire bandwidth of the switch is reserved. However, we shall sometimes consider the substochastic case in which we only have  $\sum_i \rho_{ij} \leq 1$  and  $\sum_j \rho_{ij} \leq 1$ . In this case the residual bandwidth of the switch could be used by best effort traffic.

By standard results of Birkhoff and von Neumann (see e.g. [1]) we can decompose the matrix  $M$  into a convex combination of permutation matrices,

$$M = \sum_{k=1}^K \varphi_k M_k,$$

where  $K \leq I^2 - 2I + 2$ . Here,  $M_k$  is a permutation matrix (a 0–1 matrix with exactly one “1” in each row and column),  $\varphi_k$  is the *rate* of matrix  $M_k$  and  $\sum_{k=1}^K \varphi_k = 1$ . Let  $S_{ij}$  be the set of matrices in the decomposition that have a 1 in the  $ij$  position. Then  $\rho_{ij} = \sum_{k \in S_{ij}} \varphi_k$ . Our aim is to create a schedule in which exactly one of the permutation matrices is scheduled in each time slot. Input-output pair  $ij$  is served whenever a matrix from the set  $S_{ij}$  is scheduled. Hence we require that a matrix from  $S_{ij}$  is scheduled approximately once every  $1/\rho_{ij}$  time slots.

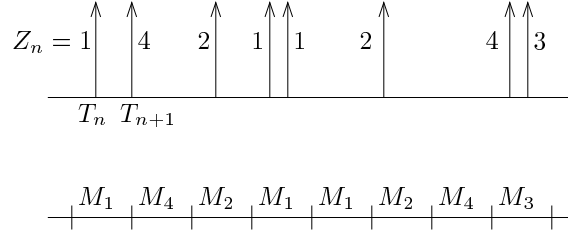


Figure 1: (Top) The token process.  $T_n$  is the time at which the  $n$ th token appears.  $Z_n$  is the type of the  $n$ th token, i.e. if the  $n$ th token corresponds to permutation matrix  $M_k$  then  $Z_n = k$ . (Bottom) The corresponding schedule.

As an example, suppose that,

$$M = \begin{pmatrix} 1/6 & 5/6 & 0 \\ 1/2 & 1/6 & 1/3 \\ 1/3 & 0 & 2/3 \end{pmatrix}.$$

Then,

$$M = \frac{1}{2}M_1 + \frac{1}{3}M_2 + \frac{1}{6}M_3,$$

where,

$$M_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, M_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, M_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and a possible schedule is,

$$M_1, M_2, M_1, M_2, M_1, M_3, M_1, M_2, M_1, M_2, M_1, M_3, \dots$$

The class of the schedulers that we consider can be formulated by the following unifying framework. We first place *tokens* for each matrix  $M_k$  in continuous time. We schedule matrix  $M_k$  in time slot  $n$  if the  $n$ th token to appear corresponds to matrix  $M_k$ . (See Figure 1).

More formally, we associate with  $M_k$  a *counting process*  $N_k$  defined on  $\mathbb{R}^+$ . For any interval  $\mathcal{I} \subseteq \mathbb{R}^+$ ,  $N_k \mathcal{I}$  equals the number of tokens for  $M_k$  that land in interval  $\mathcal{I}$ . We require that  $N_k$  has *intensity*  $\varphi_k$ , i.e.  $\lim_{t \rightarrow \infty} N_k[0, t)/t = \varphi_k$ .

We define the superposition process  $N\mathcal{I} = \sum_{k=1}^K N_k \mathcal{I}$  to which there is an associated point process  $(T_n)_{n \geq 0}$  defined on  $\mathbb{R}^+$ . Next, let  $(Z_n)_{n \geq 0}$  be the sequence of marks such that  $Z_n = k$  if and only if the  $n$ -th point of the superposition point process,  $T_n$ , belongs to  $N_k$ . Notice that we do not assume that  $(T_n)_n$  is a simple point process; if the points coincide then the ties are broken according to some rule. Let  $N_{ij}\mathcal{I}$  be the number of tokens for input-output pair  $ij$  that land in the interval  $\mathcal{I}$ , i.e.  $N_{ij}\mathcal{I} = \sum_{k \in S_{ij}} N_k \mathcal{I}$ . Likewise, let  $N_{\overline{ij}}\mathcal{I}$  be the number of tokens that land in  $\mathcal{I}$ , but do not belong to  $S_{ij}$ .

The schedule is given by the sequence  $(Z_n)_n$ . If for any given  $n$ ,  $Z_n = k$ , then the matrix  $M_k$  is scheduled in the  $n$ th slot. We say the  $n$ th token is of type  $k$ . Notice that by taking  $(Z_n)_n$  we in fact construct a *non-idle* schedule. A key feature of this schedule is,

**Observation 1** *The total number of slots in which input-output pair  $ij$  can be served during the time slots  $n, n+1, \dots, n+m-1$  is equal to  $N_{ij}[T_n, T_{n+m})$ .*

## Service Characterization

We give different characterizations of the service offered to an arbitrary input-output pair  $ij$ . Informally speaking, we would like  $N_{ij}[T_n, T_{n+m})$  to be close to  $\rho_{ij}m$ . The following is the simplest, but weakest, service characterization: for some fixed  $n \geq 0$  and  $m > 0$ ,  $E_1^{ij} \geq 0$ ,

$$\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_1^{ij})\}. \quad (1)$$

If the above event holds with probability  $1 - \varepsilon$ , a probabilistic interpretation of the service offered is: for fixed  $m$  one picks at random a slot  $n$ , then, the number of slots given to the input-output pair  $ij$  in the next  $m$  slots is at least  $\rho_{ij}(m - E_1^{ij})$  with probability  $1 - \varepsilon$ .

A natural extension of the above characterization is by requiring for some fixed  $n \geq 0$  and  $E_2^{ij} \geq 0$ ,

$$\{\forall m > 0 : N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_2^{ij})\}. \quad (2)$$

The strongest guarantee is offered by requiring, for some fixed  $E_3^{ij} \geq 0$ ,

$$\{\forall n \geq 0 \forall m > 0 : N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_3^{ij})\}. \quad (3)$$

If this event holds then we lower bound the service offered to input-output pair  $ij$  over *any* interval of time slots. In particular, it can be seen that  $f(m) = \rho_{ij}(m - E_3^{ij})$  is a strict minimum service curve offered to the  $ij$ th pair (see Proposition 1.3.6, Section 1.3.2, [15]). The service curve is “rate-latency” with rate  $\rho_{ij}$  and latency  $E_3^{ij}$ .

The service characterizations introduced so far bound how much the service offered is behind the service that would be offered by an idealistic fluid system (which would serve  $\rho_{ij}m$  bits in  $m$  slots). Thus, these service characterizations bound the *lateness* of the scheduler. Analogous characterizations can be established to bound the *earliness*; one only needs to reverse the inequalities in the above definitions, and replace minus with plus in the rate-latency functions. Small earliness of the schedule is desirable to reduce burstiness at the output of the switch.

## What Can We Compute from the Service Characterizations?

Consider an arbitrary input-output port pair  $ij$ . Let  $A_{ij}[n]$  be number of the bits that arrive in  $[0, n]$  at input port  $i$  and are destined for output port  $j$ . Then, by the result known for variable-capacity nodes

(see [15], Sec. 1.3.2, also Sec. 4.3.2), we know that the number of bits in  $[0, n]$  observed at the output port  $j$  that originate from  $i$ ,  $A_{ij}^*[n]$ , satisfies,

$$A_{ij}^*[n] = \min_{1 \leq m \leq n} [A_{ij}[m] + N_{ij}[T_m, T_n)].$$

In particular, suppose that the arrivals are  $(\sigma_{ij}, \rho_{ij})$ -*bounded*, i.e.,  $A_{ij}[n] - A_{ij}[m] \leq \rho_{ij}(n - m) + \sigma_{ij}$  for all  $m \leq n$ . Then, the following is a classical network calculus result.

**Fact 2** *The backlog of  $ij$  packets waiting for service at the switch is at most  $\sigma_{ij} + \rho_{ij}E_3^{ij}$ . If FIFO scheduling is used within the aggregate of  $ij$  packets then the maximum delay for these packets is at most  $\sigma_{ij}/\rho_{ij} + E_3^{ij}$ .*

If  $\sigma_{ij} = 0$  (i.e. the arrivals are bounded by a idealized fluid system of rate  $\rho_{ij}$ ) then the packet delay is bounded by  $E_3^{ij}$  (assuming FIFO scheduling within the aggregate). However, in a perfect schedule for  $ij$ , a matrix in  $S_{ij}$  would appear *exactly once* every  $1/\rho_{ij}$  time slots. In this case the packet delay would be  $1/\rho_{ij}$ . Hence, if  $\sigma_{ij} = 0$  we have,

$$\frac{\text{worst case packet delay}}{\text{optimal packet delay}} \leq \rho_{ij}E_3^{ij}.$$

For these reasons our objective is to keep  $E_3^{ij}$  small.

## Algorithms and results

Our algorithms will be divided into two types, *frame-based* and *non-frame-based*. Suppose that for some fixed integers  $\ell_k$  and  $L$ ,  $\varphi_k = \ell_k/L$ . (We note that this is always possible if the  $\varphi_k$  are rational.) We can compute a schedule for the interval  $[0, L)$  that contains exactly  $\ell_k$  occurrences of the permutation matrix  $m_k$  and then simply repeat this schedule for all subsequent intervals of length  $L$ . We call such a schedule a *frame-based* schedule of length  $L$ . Notice that the frame length  $L$  and number of the permutation matrices  $K$  are related as  $K = L/\bar{\ell}$ , where  $\bar{\ell}$  is the arithmetic mean of  $\ell_k$ ,  $k = 1, \dots, K$ . Typically,  $L$  would be larger than  $K$ .

If the schedule is not periodic in the above way then we say that it is *non-frame-based*. For a non-frame-based schedule we have to define it explicitly in the entire interval  $[0, \infty)$ .

In [1], Chang et al. propose a non-frame-based algorithm in which the permutation matrices are scheduled according to a PGPS [13] system that is fed with its own departures (which is initialized such that all tokens arrive at time 0). In our setting, this corresponds to placing the  $n$ th token for matrix  $M_k$  at time  $n/\varphi_k$ . More formally, for each  $k = 1, \dots, K$ ,

$$N_k[0, t) = \sum_{n \geq 0} \mathbf{1}_{[0, t)}\left(\frac{n}{\varphi_k}\right).$$

Chang et al. show that for this algorithm,

$$E_3^{ij} \leq \min\left[\frac{K}{\rho_{ij}}, \frac{|S_{ij}|}{\rho_{ij}} + (K - 1)\right]. \quad (4)$$

The aim of our work is to show that by using probabilistic techniques, it is possible to tighten this bound in many scenarios.

Our results are as follows.

1. We begin in Section 3.1 with an extremely simple frame-based scheduler in which the tokens for the permutation matrices in a frame are randomly permuted. We call this the *Random Permutation* scheduler. We require that (3) holds with probability  $1 - \varepsilon$  and we show that as  $L \rightarrow \infty$ , the latency  $E_3^{ij}$  satisfies,

$$E_3^{ij} \rightarrow \sqrt{A \left( \frac{1}{\rho_{ij}} - 1 \right) L}, \quad (5)$$

where  $A$  is a constant depending on  $\varepsilon$ . For  $E_2^{ij}$  the same expression holds, with  $A = \frac{1}{2} \ln \frac{1}{\varepsilon}$ .

2. In Section 3.2 we present a *deterministic* frame-based algorithm. We require that (3) holds with probability 1 and we show that,

$$E_3^{ij} \leq \frac{|S_{ij}|}{\rho_{ij}} + (2 + \sqrt{2K \ln(2L + 1)}). \quad (6)$$

We derive this algorithm from a randomized algorithm in which the  $n$ th token for matrix  $M_k$  is placed at time  $U_k + n/\varphi_k$  where  $U_k$  is chosen uniformly at random from  $[0, 1/\varphi_k)$ . We call this the *Random-Phase Periodic Competition* scheduler. We then show how to *derandomize* this scheduler to obtain a deterministic algorithm using the method of conditional probabilities [16]. In Section 4 we show that in many scenarios, (6) is significantly smaller than (4), largely due to the presence of the square-root in (6).

3. In Section 3.3 we present a deterministic non-frame-based algorithm. We require that (3) holds with probability 1 and we show that,

$$E_3^{ij} \leq \frac{1}{\rho_{ij}} \sqrt{2|S_{ij}| \ln D} + (2 + \sqrt{2K \ln D}), \quad (7)$$

where  $D = 1 + (4(2I^2 + 2)/\min_k \varphi_k)$ . This algorithm is derived from a randomized algorithm in which the  $n$ th token for matrix  $M_k$  is placed uniformly at random in the interval  $[(n - 1)/\varphi_k, n/\varphi_k)$ . We call this the *Random-Distortion Periodic Competition* scheduler. By using the method of conditional probabilities we are able to *derandomize* this scheduler although the analysis is more complex than it was for the random-phase scheduler since we now have to consider the entire interval  $[0, \infty)$  instead of a finite frame. In Section 4 we show that in many scenarios, (7) is significantly smaller than (4), largely due to the presence of the square-roots in (7).

4. In Section 3.4 we analyze a non-frame-based scheduler in which the tokens for matrix  $M_k$  are placed according to a Poisson process. We call this the *Poisson Competition* scheduler. For this scheduler only we assume that the load on each input and output is strictly less than 1. We show using the Brownian approximation [17] that,

$$E_2^{ij} \approx \frac{1}{2} \ln\left(\frac{1}{\varepsilon}\right) \frac{\rho}{1-\rho} \left( \frac{1-\rho}{\rho_{ij}} + \rho \right), \quad (8)$$

where  $\rho$  is the load. The latency  $E_3^{ij}$  does not make much sense for this scheduler since the event in (3) would fail with probability one as we require that the inequality in (3) holds for all  $n$ .

## Comparison with single server polling

We remark that our problem is significantly different from the single server polling problem (e.g. see [18] and references therein) in which a single server has to poll a set of clients at predetermined frequencies. Note that in our problem it is not sufficient for each matrix  $M_k$  to be served at evenly spaced intervals of  $1/\varphi_k$  slots. This is because input-output pair  $ij$  is served whenever a matrix in  $S_{ij}$  is served. If  $k, \ell \in S_{ij}$  and  $M_k$  and  $M_\ell$  are served close together then the service to  $ij$  is bursty even though each permutation matrix might receive smooth service. Note however that we cannot in isolation change the schedule to improve service for one particular input-output pair since each permutation matrix is a member of  $S_{ij}$  for  $I$  different pairs  $ij$ .

## Previous work

As mentioned earlier, papers that analyze schedulers based on decomposing the rate matrix include [1, 2, 3]. Analyses of MWM-type schedulers can be found in, for example, [7, 8, 9, 10, 11]. Frame-based schedulers were presented in [19, 20]. If the switch fabric has an internal speedup of 2 then it is known that it can emulate output-queued switches (in which there is no contention at the inputs) [21, 22, 23]. In [24], an algorithm is presented whose aim is to “track” an idealized fluid policy.

If the switch is sufficiently underloaded then tight delay bounds can be achieved. In [25] it is shown that if the total load on any input or output is at most one quarter of the link rate, then it is possible to serve each  $ij$  pair at least once every  $1/\rho_{ij}$  steps.

The remainder of the paper is organized as follows. In Section 2 we derive some useful facts about the event,  $\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E)\}$ . In Section 3 we present our four schedulers and analyze them in detail. In Section 4 we present some numerical results to evaluate our bounds for specific rate matrices. We defer proofs of some of our results to the Appendices.



## 2 Preliminary Analysis

Observe that the event (1) is equivalent to:

$$\{\exists_{t>0} : N_{ij}[T_n, T_n + t) \geq \rho_{ij}(m - E_1^{ij}), N[0, T_n + t) = n + m - 1\}.$$

Since there is a point at the left end of the interval  $[T_n, T_n + t)$ , the probability of the above event is with respect to Palm probability [26]. This amounts to the equivalent representation,

$$\{\exists_{t>0} \exists_{s \geq 0} : N_{ij}[s, s + t) \geq \rho_{ij}(m - E_1^{ij}), N[0, s + t) = n + m - 1, N[0, s) = n - 1\}.$$

Unfortunately, it is hard in general to calculate the probability of the above event since there is too much dependence between the constituent events. It is however feasible for the case of point processes with independent increments. An example of this special case is the Poisson Competition scheduler that we analyze in Section 3.4 using a *Geo/D/1* queue.

In the remainder of the section, we try to define a subevent of (1) whose probability is easier to bound in the general case. To that end, let  $G_{n,m}$  be the *good* event  $\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E)\}$ . Let  $B_{n,m}$  be the *bad* event  $B_{n,m} = \overline{G_{n,m}}$ .

Let  $\Delta_1, \Delta_2 \in \mathbb{Z}^+$  and  $\Delta_3, \Delta_4 \in \mathbb{R}^+$  satisfy

$$\Delta_1 + \Delta_2 + (\Delta_3 + \Delta_4)/\rho_{ij} \leq E,$$

where  $E = E_1^{ij}, E_2^{ij}$  or  $E_3^{ij}$ , depending on our calculation.

Let  $t = n + m - \Delta_1$  and let  $s = n + \Delta_2$ . Note that  $s$  and  $t$  are integers.

**Lemma 3** *Suppose that  $N_{ij}[s, t) \geq \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)$  and  $[s, t) \subseteq [T_n, T_{n+m})$ . Then,  $N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E)$ .<sup>1</sup>*

**Proof:** We have,

$$\begin{aligned} N_{ij}[T_n, T_{n+m}) &\geq N_{ij}[s, t) \\ &\geq \rho_{ij}(t - s) - (\Delta_3 + \Delta_4) \\ &= \rho_{ij}((n + m - \Delta_1) - (n + \Delta_2) - (\Delta_3 + \Delta_4)/\rho_{ij}) \\ &= \rho_{ij}(m - (\Delta_1 + \Delta_2) - (\Delta_3 + \Delta_4)/\rho_{ij}) \\ &\geq \rho_{ij}(m - E). \end{aligned}$$

The first two inequalities come from the assumptions of the lemma. The first equality comes from the definitions of  $s$  and  $t$ . The final inequality comes from our constraint on the  $\Delta$ 's.  $\square$

---

<sup>1</sup>Note that we are interested to the values of  $m$  such that  $m \geq E$ . This implies  $m \geq \Delta_1 + \Delta_2$ , which is equivalent to  $s \leq t$ . For  $m < E$  the inequalities in (1), (2), (3) do indeed hold.

From the definition of  $G_{n,m}$ , Lemma 3 implies

$$G_{n,m} \supseteq \{N_{ij}[s, t] \geq \rho_{ij}(t - s) - (\Delta_3 + \Delta_4)\} \cap \{[s, t] \subseteq [T_n, T_{n+m}]\}. \quad (9)$$

By the above results we can focus on the quantity  $N_{ij}[s, t]$  and the relationship between the intervals  $[s, t]$  and  $[T_n, T_{n+m}]$ . However, for the random processes we consider, each interval  $[s, t]$  will be dependent on too many other intervals. The way to solve this problem is to concentrate on intervals that have one of their endpoints fixed. For this purpose we must refine our results.

**Lemma 4** *If  $N[0, t] < t + \Delta_1$  and  $N[0, s] \geq s - \Delta_2$  then  $[s, t] \subseteq [T_n, T_{n+m}]$ .*

**Proof:** Since  $t = n + m - \Delta_1$ ,  $N[0, t] < t + \Delta_1 \Rightarrow [0, t] \subseteq [0, T_{n+m}]$ . Similarly, since  $s = n + \Delta_2$ ,  $N[0, s] \geq s - \Delta_2 \Rightarrow [0, s] \supseteq [0, T_n]$ . Therefore,  $[s, t] \subseteq [0, T_{n+m}] \setminus [0, T_n) = [T_n, T_{n+m}]$ .  $\square$

**Lemma 5** *If  $N_{ij}[0, t] \geq \rho_{ij}t - \Delta_3$  and  $N_{ij}[0, s] \leq \rho_{ij}s + \Delta_4$  then  $N_{ij}[s, t] \geq \rho_{ij}(t - s) - (\Delta_3 + \Delta_4)$ .*

**Proof:** We have,  $N_{ij}[s, t] = N_{ij}[0, t] - N_{ij}[0, s] \geq (\rho_{ij}t - \Delta_3) - (\rho_{ij}s + \Delta_4) = \rho_{ij}(t - s) - (\Delta_3 + \Delta_4)$ .  $\square$

Lemma 4, Lemma 5 and (9) imply,

$$G_{n,m} \supseteq \{N[0, t] < t + \Delta_1\} \cap \{N[0, s] \geq s - \Delta_2\} \cap \{N_{ij}[s, t] \geq \rho_{ij}(t - s) - (\Delta_3 + \Delta_4)\}, \quad (10)$$

and,

$$G_{n,m} \supseteq \{N[0, t] < t + \Delta_1\} \cap \{N[0, s] \geq s - \Delta_2\} \cap \{N_{ij}[0, t] \geq \rho_{ij}t - \Delta_3\} \cap \{N_{ij}[0, s] \leq \rho_{ij}s + \Delta_4\}. \quad (11)$$

If we are interested in calculating  $E_1^{ij}$  then we only need to focus on some fixed  $n$  and  $m$ .

However, if we are interested in  $E_3^{ij}$  then we need to know whether  $G_{n,m}$  for *all*  $n, m$ . For the latter case we have,

$$\bigcap_{n,m} G_{n,m} \supseteq \bigcap_t \{N[0, t] < t + \Delta_1\} \cap \bigcap_s \{N[0, s] \geq s - \Delta_2\} \cap \bigcap_{s,t} \{N_{ij}[s, t] \geq \rho_{ij}(t - s) - (\Delta_3 + \Delta_4)\}. \quad (12)$$

and,

$$\bigcap_{n,m} G_{n,m} \supseteq \bigcap_t \{N[0, t] < t + \Delta_1\} \cap \bigcap_s \{N[0, s] \geq s - \Delta_2\} \cap \bigcap_t \{N_{ij}[0, t] \geq \rho_{ij}t - \Delta_3\} \cap \bigcap_s \{N_{ij}[0, s] \leq \rho_{ij}s + \Delta_4\}. \quad (13)$$

We note that since  $s$  and  $t$  are integers we only need to take the intersection over a discrete set of events.

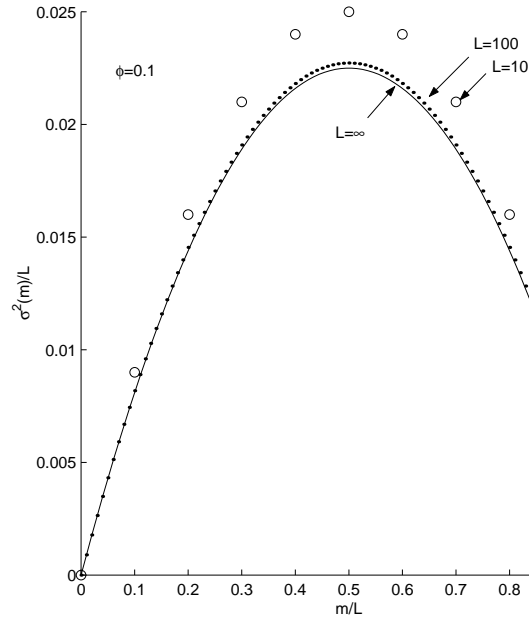


Figure 2: Normalized variance of  $N_{ij}[T_n, T_{n+m}]$  for varying frame size  $L$ , and  $\rho_{ij} = \phi = 1/10$ .

### 3 Four Schedulers

#### 3.1 Random Permutation

We consider a frame-based scheduler in which the permutation matrices in a frame are scheduled in random order. More formally, denote by  $z = [z_1, z_2, \dots, z_L]$  some fixed order of the token types, where there are exactly  $\ell_k$  tokens of type  $k$ ,  $k = 1, \dots, K$ . Let  $\pi = [\pi(1), \pi(2), \dots, \pi(L)]$  be a random permutation of the elements  $[1, 2, \dots, L]$ .

For  $n = 1, \dots, L$  we define the schedule by randomly permuting the elements of  $z$ , i.e.,

$$Z_n = z_{\pi(n)}, \text{ for } n = 1, \dots, L.$$

The schedule is extended for  $n > L$  by concatenating replicas of the schedule  $Z_n$ ,  $n = 1, \dots, L$ .

Notice that the scheduler as defined above fits in the framework of point processes. We can first construct the counting processes  $N_k$ ,  $k = 1, \dots, K$ , on  $[0, 1]$  by placing  $\ell_k$  points uniformly in  $[0, 1]$ . Then,  $N_k$  is extended to the whole positive real line by periodic extension of the points in  $[0, 1]$ .

We first give some elementary properties of the scheduler as described above, and then derive the latencies. By a routine combinatorial argument we obtain, for  $l = 1, \dots, \min[\ell_{ij}, m]$ ,

$$P(N_{ij}[T_n, T_{n+m}] = l) = \frac{\binom{m}{l} \binom{L-m}{\ell_{ij}-l}}{\binom{L}{\ell_{ij}}},$$

where  $\ell_{ij} = \sum_{k \in S_{ij}} \ell_k$ . Notice that the probability does not depend on  $n$ , which reflects the fact that the process is stationary. The above explicit expression would enable us to compute the latency  $E_1^{ij}$  from (1). As an aside, note that for fixed  $m$ ,  $N_{ij}[T_n, T_{n+m})$ , as  $L \rightarrow \infty$ , converges in distribution to Binomial random variable  $(m, \rho_{ij})$  (this can be checked, for example, by Stirling's formula). The variance of  $N_{ij}[T_n, T_{n+m})$  is given by,

$$\sigma_{ij}^2(m) = \frac{L^2}{L-1} \rho_{ij}(1-\rho_{ij}) \frac{m}{L} \left(1 - \frac{m}{L}\right).$$

It forms a bridge ( $\sigma_{ij}^2(0) = \sigma_{ij}^2(L) = 0$ , with the global maximum on  $m \in [0, L]$ , at  $m = L/2$ ); see Figure 2. Note also that,  $\sigma_{ij}^2(m) \rightarrow L \rho_{ij}(1-\rho_{ij}) \frac{m}{L} \left(1 - \frac{m}{L}\right)$ , as  $L \rightarrow \infty$ .

We obtain simple asymptotic expressions for the latencies  $E_2^{ij}$  and  $E_3^{ij}$ . For reasons of space, the proofs are deferred to the Appendices.

**Proposition 6** As  $L \rightarrow \infty$ ,

$$E_2^{ij} \rightarrow \sqrt{\frac{1}{2} \ln \frac{1}{\varepsilon} \left(\frac{1}{\rho_{ij}} - 1\right) L}.$$

It is worth mentioning that the latency scales with the frame length  $L$  as  $O(\sqrt{L})$ .

We have the following result for the latency  $E_3^{ij}$ .

**Proposition 7** As  $L \rightarrow \infty$ ,

$$E_3^{ij} \rightarrow \sqrt{A \left(\frac{1}{\rho_{ij}} - 1\right) L}, \tag{14}$$

where  $A$  is the positive solution of  $\sum_{\ell \geq 1} (4\ell^2 A - 1) e^{-2\ell^2 A} = \varepsilon$ .

Note that the latency  $E_3^{ij}$  also scales with the frame length  $L$  as  $O(\sqrt{L})$ , but with a different constant. In Figure 3 we plot values for  $E_2^{ij}$  and  $E_3^{ij}$  determined empirically together with the above limits for different values of  $L$ .

### 3.2 Random-Phase Periodic Competition

Let  $U_k, k = 1, \dots, K$ , be a collection of independent uniformly distributed random variables on  $[0, 1]$ . We define the scheduler as follows; for each  $k = 1, \dots, K$ ,

$$N_k[0, t) = \sum_{n \geq 0} \mathbf{1}_{[0, t)} \left( \frac{n-1}{\varphi_k} + \frac{1}{\varphi_k} U_k \right).$$

Thus the tokens for matrix  $M_k$  form a periodic stream of period  $1/\varphi_k$  with the random phase shift  $U_k/\varphi_k$ .

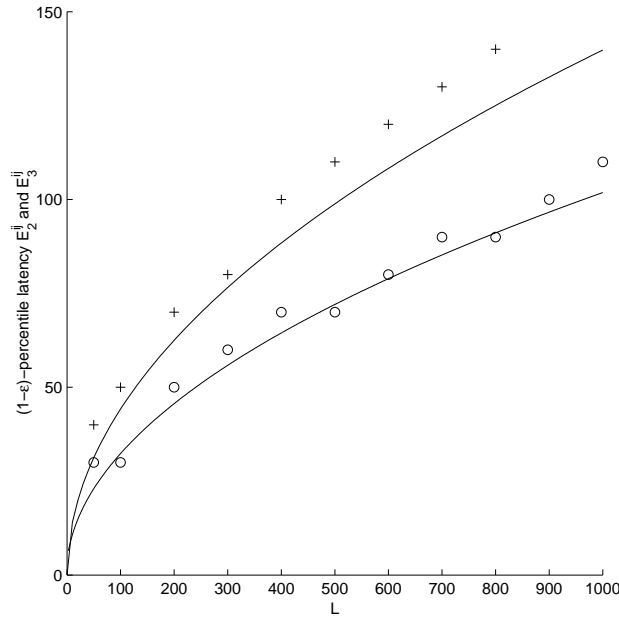


Figure 3: (Circles and bottom curve) Empirical values and limit curve for  $E_2^{ij}$ . (Crosses and top curve) Empirical values and limit curve for  $E_3^{ij}$ .  $\varepsilon = 0.1$ .

We assume that the  $\varphi_k$  have the property that we can define a frame-based scheduler (see the Introduction). Therefore we only need to concentrate on the time interval  $L$ . For any interval  $[s, t]$ ,  $N_k[s, t] \geq \varphi_k(t - s) - 1$ . This implies,

$$N_{ij}[s, t] \geq \sum_{k \in S_{ij}} (\varphi_k(t - s) - 1) \Rightarrow N_{ij}[s, t] \geq \rho_{ij}(t - s) - |S_{ij}|.$$

We follow the method of Section 2 and set  $t = n + m - \Delta_1$ ,  $s = n + \Delta_2$ , where  $\Delta_1, \Delta_2$  are defined below. For any permutation matrix  $M_k$ , let  $a_k = \lfloor \varphi_k t \rfloor$ . Then  $N_k[0, t] = a_k + X_k$  where  $X_k$  is a binary random variable with mean  $\varphi_k t - a_k$ . Let  $\mu = \mathbb{E}[\sum_k X_k] = \sum_k (\varphi_k t - a_k) = t - \sum_k a_k$ . We have,

$$\begin{aligned} N[0, t] < t + \Delta_1 &\Leftrightarrow \sum_k (a_k + X_k) < t + \Delta_1 \\ &\Leftrightarrow \sum_k X_k < t + \Delta_1 - \sum_k a_k \\ &\Leftrightarrow \sum_k X_k < \mu + \Delta_1. \end{aligned}$$

Therefore, by Hoeffding's inequality [27],  $\mathbb{P}(N[0, t] \geq t + \Delta_1) \leq \exp(-2(\Delta_1)^2/K)$ . Similarly,  $\mathbb{P}(N[0, s] < s - \Delta_2) \leq \exp(-2(\Delta_2)^2/K)$ . Let,  $\Delta_1 = \Delta_2 = \left\lceil \sqrt{(K/2) \cdot \ln(2L + 1)} \right\rceil$ , let  $\Delta_3 = \Delta_4 = |S_{ij}|/2$  and let,

$$E_3^{ij} = (|S_{ij}|/\rho_{ij}) + (2 + \sqrt{2K \ln(2L + 1)}).$$

Then, by the above Hoeffding bounds and the containment (12) from Section 2 we have,

$$\begin{aligned}
& \mathbb{P}(\bigcap_{ij} \bigcap_{nm} \{N_{ij}[T_n, T_m] \geq \rho_{ij}(m - E_3^{ij})\}) \\
& \geq 1 - \sum_{t=1}^L \mathbb{P}(N[0, t] \geq t + \Delta_1) - \sum_{s=1}^L \mathbb{P}(N[0, s] < s - \Delta_2) \\
& \geq 1 - \sum_{t=1}^L \exp\left(\frac{-2(\Delta_1)^2}{K}\right) - \sum_{s=1}^L \exp\left(\frac{-2(\Delta_2)^2}{K}\right) \\
& \geq 1 - \frac{2L}{2L+1}.
\end{aligned} \tag{15}$$

(Note that since we are considering a finite frame we only need sum over  $s, t \in \{1, \dots, L\}$ .)

Hence with probability  $1 - 2L/(2L + 1)$ , the rate-latency condition (3) holds *for all*  $ij$ . We now show how to *derandomize* the algorithm so that condition (3) holds with probability 1.

## Derandomization

We use the method of conditional probabilities (see e.g. [16]) that is motivated by the following lemma (which we prove in Appendix C).

**Lemma 8** *Let  $Y_1, \dots, Y_{n_1}$  be a set of random variables, let  $X_1, \dots, X_{n_2}$  be a set of independent binary random variables and let  $\sigma_1, \dots, \sigma_{n_3}$  be a set of events such that for some functions  $f_{ij}(\cdot)$ ,*

$$\mathbb{P}(\sigma_i | A) \leq \mathbb{E}\left[\prod_{j=1}^{n_2} f_{ij}(X_j) | A\right], \tag{16}$$

*for any event  $A$ . Then there exists a set of values  $y_1, \dots, y_{n_1}$  such that,*

$$\sum_i \mathbb{P}(\sigma_i | Y_1 = y_1, \dots, Y_{n_1} = y_{n_1}) \leq \sum_i \mathbb{E}\left[\prod_{j=1}^{n_2} f_{ij}(X_j)\right].$$

*In particular if  $\sum_i \mathbb{E}\left[\prod_{j=1}^{n_2} f_{ij}(X_j)\right] < 1$  and  $\sigma_i$  is completely determined by  $Y_1, \dots, Y_{n_1}$  then,*

$$\sum_i \mathbb{P}(\sigma_i | Y_1 = y_1, \dots, Y_{n_1} = y_{n_1}) = 0.$$

To compute  $y_v$  given  $y_1, \dots, y_{v-1}$  we minimize,

$$\sum_i \mathbb{E}\left[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_v = y_v\right] = \sum_i \prod_{j=1}^{n_2} \mathbb{E}[f_{ij}(X_j) | Y_1 = y_1, \dots, Y_v = y_v],$$

as we vary  $y_v$  over the full range of  $Y_v$ . (Recall that the random variables  $X_j$  are independent and so we can exchange the expectation with the product).

To apply this lemma in our setting we take  $Y_1, \dots, Y_{n_1}$  to be the random phase shifts  $U_1, \dots, U_K$ ;  $X_1, \dots, X_{n_2}$  to be binary random variables of the form  $N_k[0, t] - \lfloor \varphi_k t \rfloor$  and  $\sigma_1, \dots, \sigma_{n_3}$  to be events of the form  $\{N[0, t] \geq t + \Delta_1\}$  or  $\{N[0, s] < s - \Delta_2\}$ . The functions  $f_{ij}(\cdot)$  are defined by,

$$\mathbb{P}(N[0, t] \geq t + \Delta_1) \leq e^{-\theta(t+\Delta_1)} \mathbb{E}\left[\prod_{k=1}^K e^{\theta N_k[0, t]}\right], \tag{17}$$

where  $\theta = \ln(b(1-a)/a(1-b))$ ,  $a = (t - \sum_{k=1}^K \lfloor \varphi_k t \rfloor)/K$ ,  $b = (t + \Delta_1 - \sum_{k=1}^K \lfloor \varphi_k t \rfloor)/K$ . A similar inequality holds for  $P(N[0, s] < s - \Delta_2)$ .

In the derivation of (15) we showed that,

$$\sum_{t=1}^L P(N[0, t] \geq t + \Delta_1) + \sum_{s=1}^L P(N[0, s] < s - \Delta_2) \leq \frac{2L}{2L+1},$$

using Hoeffding bounds that are derived from (17). Hence by Lemma 8 there exist fixed values  $u_1, \dots, u_K$  for the initial phase shifts such that  $N[0, t] < t + \Delta_1$  for all  $t$  and  $N[0, s] \geq s - \Delta_2$  for all  $s$ .

The one complication that arises in the calculation of the  $u_k$  is that the  $U_k$  are continuous random variables, they do not take discrete values. However, as  $U_k$  is varied between 0 and 1,  $N_k[0, t] - \lfloor \varphi_k t \rfloor$  changes from 0 to 1 at one discrete point. Hence it is sufficient to consider only  $L+1$  values of  $U_k$ . The right-hand side of (17) may be computed in time polynomial in  $K$  and  $L$ , even if some of the phase shifts have already been fixed. Hence, we can fix the value of  $U_k$  in time polynomial in  $K$  and  $L$ .

**Theorem 9** *The resulting deterministic scheduler satisfies (3) with,*

$$E_3^{ij} = \frac{|S_{ij}|}{\rho_{ij}} + (2 + \sqrt{2K \ln(2L+1)}).$$

### 3.3 Random-Distortion Periodic Competition

Let  $U_{kn}$ ,  $k = 1, \dots, K$ ,  $n \in \mathbb{Z}^+$ , be a collection of independent uniformly distributed random variables on  $[0, 1]$ . We define the scheduler as follows; for each  $k = 1, \dots, K$ ,

$$N_k[0, t] = n + 1_{U_{kn} < \varphi_k t - n},$$

where  $n = \lfloor \varphi_k t \rfloor$ . Another interpretation is: the  $n$ th point of the  $k$ th token type is placed uniformly at random in the interval  $[n/\varphi_k, (n+1)/\varphi_k)$ .

We make use of the containment (13) from Section 2. We apply Hoeffding bounds in a similar manner to the previous subsection to obtain.

$$\begin{aligned} P(N[0, t] \geq t + \Delta_1) &\leq \exp(-2(\Delta_1)^2/K), & P(N[0, s] < s - \Delta_2) &\leq \exp(-2(\Delta_2)^2/K), \\ P(N_{ij}[0, t] \leq \rho_{ij}t - \Delta_3) &\leq \exp(-2(\Delta_3)^2/|S_{ij}|), & P(N_{ij}[0, s] \geq \rho_{ij}s + \Delta_4) &\leq \exp(-2(\Delta_4)^2/|S_{ij}|). \end{aligned}$$

Let,

$$\begin{aligned}
\Gamma_1(t) &= \mathbb{P}(N[0, t] \geq t + \Delta_1) + \sum_{ij} \mathbb{P}(N_{ij}[0, t] < \rho_{ij}t - \Delta_3), \\
\Gamma_2(s) &= \mathbb{P}(N[0, s] < s - \Delta_2) + \sum_{ij} \mathbb{P}(N_{ij}[0, s] > \rho_{ij}s + \Delta_4), \\
D &= 1 + (4(2I^2 + 2)/\min_k \varphi_k), \\
\Delta_1 = \Delta_2 &= \left\lceil \sqrt{(K/2) \cdot \ln D} \right\rceil, \\
\Delta_3 = \Delta_4 &= \sqrt{(|S_{ij}|/2) \cdot \ln D}, \\
E_3^{ij} &= \Delta_1 + \Delta_2 + (\Delta_3 + \Delta_4)/\rho_{ij}.
\end{aligned}$$

For fixed  $s$  and  $t$  we have,

$$\Gamma_1(t) + \Gamma_2(s) \leq \frac{(2I^2 + 2)}{D}.$$

Note that we cannot apply a union bound over  $s$  and  $t$  as we did in the previous subsection because  $s$  and  $t$  range over the entire interval  $[0, \infty)$ . However, note that if  $\Gamma_1(t) = \Gamma_2(s) = 0$  for all  $s, t$  then from (13) we know that (3) holds for all  $i, j$  with probability 1. Hence we focus on derandomizing the algorithm.

## Derandomization

Instead of placing the  $n$ th token for matrix  $M_k$  at random into the interval  $[n/\varphi_k, (n+1)/\varphi_k)$ , we now wish to place it deterministically. Let  $P = \lceil 2/\min_k \varphi_k \rceil$ . We divide time into intervals of length  $P$ , namely,  $[0, P), [P, 2P), \dots$ . Let  $A^\omega$  be the set of tokens that fall into the interval  $[\omega P, (\omega+1)P)$  with probability 1, i.e. the  $n$ th token for matrix  $M_k$  is in  $A^\omega$  if and only if  $[n/\varphi_k, (n+1)/\varphi_k) \subseteq [\omega P, (\omega+1)P)$ . Let  $B^\omega$  be the set of tokens that are not in  $A^{\omega'}$  for any  $\omega'$  and that fall into the interval  $[(\omega - \frac{1}{2})P, (\omega + \frac{1}{2})P)$  with probability 1. We have chosen  $P$  sufficiently large so that all tokens are in  $A^\omega \cup B^\omega$  for some  $\omega$ .

Suppose inductively that for  $\omega' < \omega$  we have fixed the positions of all the tokens in  $A^{\omega'} \cup B^{\omega'}$ . Since none of the tokens that have already been fixed affect the interval  $[\omega P, (\omega+1)P)$ , our previous analysis implies,

$$\sum_{\omega P}^{(\omega+1)P} \Gamma_1(t) + \sum_{\omega P}^{(\omega+1)P} \Gamma_2(s) \leq \frac{P(2I^2+2)}{D}.$$

By applying the method of conditional probabilities in a similar manner to Section 3.2, we can fix the positions of tokens in  $A^\omega$  one after the other so that we still have,

$$\sum_{\omega P}^{(\omega+1)P} \Gamma_1(t) + \sum_{\omega P}^{(\omega+1)P} \Gamma_2(s) \leq \frac{P(2I^2+2)}{D}.$$

Here, the constituent probabilities of  $\Gamma_1(t)$  and  $\Gamma_2(t)$  are now conditioned on the fact that the tokens in  $A^\omega$  are fixed. Therefore,

$$\begin{aligned}
\sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_1(t) + \sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_2(s) &\leq \sum_{(\omega-1)P}^{(\omega+1)P} \Gamma_1(t) + \sum_{(\omega-1)P}^{(\omega+1)P} \Gamma_2(s) \\
&\leq \frac{2P(2I^2+2)}{D} < 1.
\end{aligned}$$



By the method of conditional probabilities we can fix the positions of tokens in  $B^\omega$  so that we still have,

$$\sum_{(\omega - \frac{1}{2})P}^{(\omega + \frac{1}{2})P} \Gamma_1(t) + \sum_{(\omega - \frac{1}{2})P}^{(\omega + \frac{1}{2})P} \Gamma_2(s) \leq \frac{2P(2I^2+2)}{D} < 1.$$

All tokens in  $A^\omega \cup B^\omega$  are now fixed and so we have a deterministic schedule up to time  $(\omega + \frac{1}{2})P$ . Recall that  $\Gamma_1(t)$  and  $\Gamma_2(s)$  are sums of probabilities. Therefore  $\Gamma_1(t) = \Gamma_2(s) = 0$  for all  $s, t \in [(\omega - \frac{1}{2})P, (\omega + \frac{1}{2})P)$ . This process can be repeated indefinitely.

**Theorem 10** *The resulting deterministic scheduler satisfies (3) with,*

$$E_3^{ij} = \frac{1}{\rho_{ij}} \sqrt{2|S_{ij}| \ln D} + (2 + \sqrt{2K \ln D}).$$

## Adaptation to the substochastic case

For the previous three schedulers, we have assumed that the rate matrix  $M$  is doubly stochastic, i.e.  $\sum_i \rho_{ij} = 1$  and  $\sum_j \rho_{ij} = 1$ . For the case in which  $M$  is only substochastic, i.e.  $\sum_i \rho_{ij} \leq 1$  and  $\sum_j \rho_{ij} \leq 1$ , it is known by a result of von Neumann (see e.g. [1]) that there exists a matrix  $M'$  with  $ij$  entry  $\rho'_{ij}$  such that  $\rho_{ij} \leq \rho'_{ij}$  for all  $ij$  and  $M'$  is doubly stochastic. In this case, we can apply all the results of this paper to the matrix  $M'$  to obtain latencies  $E_1^{ij}$ ,  $E_2^{ij}$  and  $E_3^{ij}$ . Note that the  $ij$  traffic might not be able to use all the service it is offered. In this case the residual bandwidth can be used for best-effort traffic.

For our final scheduler we require that the load on each input and output is strictly less than 1.

## 3.4 Poisson Competition

Let  $N_k$  be Poisson process with intensity  $\varphi_k$ , for all  $k = 1, \dots, K$ . Then the following holds.

**Lemma 11** *For any  $n, m$ ,*

$$\mathbb{P}(N_{ij}[T_n, T_{n+m}) = l) = \binom{m}{l} \rho_{ij}^l (1 - \rho_{ij})^{m-l}.$$

We give a pedestrian proof of the lemma based on the definition and some elementary properties of Poisson process in Appendix D. In this case,  $(T_n, Z_n)_{n \geq 0}$  is a marked point process with i.i.d. marks, where  $Z_n = k$  with probability  $\varphi_k$ . Our naming of this scheduler is inspired by the Poisson competition theorem (see Theorem 1.3, Chapter 8 in [28]).

It is instructive to observe the following queueing interpretation of the latencies defined in (2) and (3). In the sequel, assume  $\varphi_k, k = 1, \dots, K$ , are such that for all  $ij$ ,  $\sum_{k \in \bar{S}_{ij}} \varphi_k = \rho(1 - \rho_{ij})$ , for some

$\rho < 1$ . We require this assumption to ensure stability. In addition, assume that our counting processes  $N_k$  are extended to the whole real line. Then, it is not difficult to observe that (2) is equivalent to:

$$\{V_{ij}^-[0] \leq \rho_{ij} E_2^{ij}\},$$

where  $V_{ij}^-$  is the unfinished work of a constant rate server with rate  $(1 - \rho_{ij})$  and infinite buffer capacity that is fed with 0 – 1 arrivals  $N_{ij}^-$  of intensity  $\rho(1 - \rho_{ij})$  (where  $N_{ij}^-$  is taken in reverse time). The above observation is made from Reich's formula:  $V_{ij}^-[n] = \max_{m \geq 1} [N_{ij}^-[T_n, T_{n+m}) - (1 - \rho_{ij})m]$ .

In the present context, this queueing interpretation, along with the above lemma, amounts to studying the unfinished work of a *Geo/D/I* queue. At each time slot, there is an arrival with probability  $\rho(1 - \rho_{ij})$  each requiring a fixed service of  $1/(1 - \rho_{ij})$ . The distribution of the unfinished work for *Geo/D/I* queue is known in closed form [29], which in our context gives:

$$P(V_{ij}^-[0] \leq v) = \frac{1 - qD}{(1 - q)^{v+1}} \sum_{l=0}^j [q(1 - q)^{D-1}]^l (-1)^l \binom{v - (D-1)l}{l},$$

for  $jD \leq v \leq (j+1)D - 1$ ,  $j = 1, 2, \dots$ . Here,  $q := \rho(1 - \rho_{ij})$  and  $D := 1/(1 - \rho_{ij})$  is an integer. (If  $D$  is a fraction, then one may redefine  $D := \lceil 1/(1 - \rho_{ij}) \rceil$  to obtain a lower bound, provided that  $\rho D < 1$ .)

The last would enable one at least in theory to exactly compute the latency  $E_2^{ij}$  in (2). In practice one may expect numerical instability as  $\rho D \rightarrow 1$ . The following heuristic argument gives us an approximation that is numerically stable, but also gives us some insight about the latency. By appealing to the Brownian approximation (see [17], Sec. 5.7, Equation (7.16)) we claim

$$P(V_{ij}^-[0] \leq v) \approx 1 - e^{-2 \frac{1-\rho}{\rho(1-\rho(1-\rho_{ij}))} \rho_{ij} E_2^{ij}}.$$

Hence we have

$$E_2^{ij} \approx \frac{1}{2} \ln\left(\frac{1}{\varepsilon}\right) \frac{\rho}{1 - \rho} \left( \frac{1 - \rho}{\rho_{ij}} + \rho \right).$$

Another approximation can be obtained by considering an *M/D/I* queue, which is a continuous time analog to *Geo/D/I*, with a notable difference – with *Geo/D/I*, as opposed to *M/D/I*, the number of arrivals over any interval of length  $m$  is bounded by  $m$ . A simple exponential approximation is known for *M/D/I* ([30], Equation 6.1.6, Section 6.1.2). In Figure 4 we show a numerical comparison of the approximations mentioned above with their empirical companions. We observe that the above approximation for  $E_2^{ij}$  should be good in the heavy-traffic limit as  $\rho \rightarrow 1$ . It is perhaps interesting to observe that in the heavy-traffic limit  $E_2^{ij}$  becomes insensitive to  $\rho_{ij}$ .

As mentioned in the Introduction, the latency  $E_3^{ij}$  does not make much sense for this scheduler since the event in (3) would fail with probability one as we require that the inequality in (3) holds for all  $n$ .

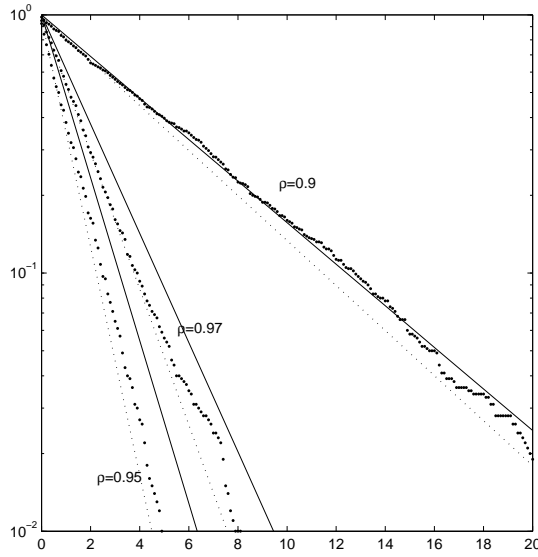


Figure 4: Tail distribution of  $V_{ij}$ : (dots) empirical estimate, (dotted line) M/D/1, (solid line) Brownian approximation.  $V_{ij}$  is estimated by averaging over 1000 random samples of length 10000.  $\rho_{ij} = 0.1$ .

## 4 Numerical Results

In this section we evaluate some of our bounds for specific rate matrices. Recall that the best possible latency for input-output pair  $ij$  is  $1/\rho_{ij}$ . Hence the ratio between the latency provided by the scheduler,  $E_3^{ij}$ , and the best possible latency is  $\rho_{ij} E_3^{ij}$ . For this reason we define  $\max_{ij} \rho_{ij} E_3^{ij}$  to be the figure of merit for a scheduler.

We evaluate the bounds (6) and (7) for the deterministic algorithms derived from Random-Phase Periodic Competition and Random-Distortion Periodic Competition. We compare them with the bound (4) for PGPS. We would like to use matrices drawn uniformly from the set of doubly-stochastic matrices. However, we do not know of a method to generate such a matrix uniformly. Hence we use the following method to generate our example matrices. We start with a uniform matrix in which all entries are equal to  $I/L$  where  $L = I \times I$ . We then repeatedly choose parameters  $i_1, i_2, j_1, j_2$  and  $\delta$  uniformly at random such that  $\delta \leq \min\{\rho_{i_1 j_1}, \rho_{i_2 j_2}\}$ . We subtract  $\delta$  from  $\rho_{i_1 j_1}$  and  $\rho_{i_2 j_2}$  and we add  $\delta$  to  $\rho_{i_1 j_2}$  and  $\rho_{i_2 j_1}$ . We carry out this operation 100000 times. Note that it preserves the doubly stochastic nature of the matrix. We also ensure that all entries of the rate matrix are integer multiples of  $1/L$ . Hence we can define frame-based schedulers with frame-length  $L$ .

In Figure 5 we plot the value of  $\max_{ij} \rho_{ij} E_3^{ij}$  for different values of  $I$ , the switch size. We see that except for extremely small switches, the bound for the Random-Distortion scheduler is smaller than the bound for the Random-Phase scheduler which is in turn smaller than the bound for PGPS.

In Figure 6 we examine how  $\rho_{ij} E_3^{ij}$  varies for different pairs  $ij$ . In particular we examine a  $64 \times 64$  matrix for which  $K = 2423$ . For each value of  $x$  we plot the fraction of  $ij$  pairs for which  $\rho_{ij} E_3^{ij} \leq x$ .

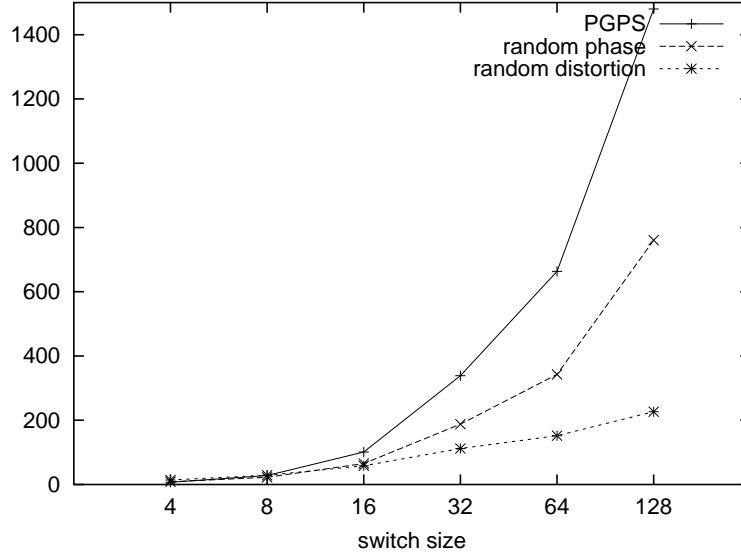


Figure 5: The value of  $\max_{ij} \rho_{ij} E_3^{ij}$  for switches of varying size.

We see that the bound (6) for the Random-Phase based algorithm is consistently smaller than the bound (4) for PGPS. The bound (7) for the Random-Distortion based algorithm has a smaller range than the other two bounds. There are fewer pairs  $ij$  with large values of  $\rho_{ij} E_3^{ij}$  but there are also fewer pairs  $ij$  with small values of  $\rho_{ij} E_3^{ij}$ . The reason for the latter phenomenon is that the bound (7) is typically larger than the bounds, (4), (6) when the value of  $|S_{ij}|$  is small.

We remark that we cannot directly compare the expressions (5) and (8) for the Random Permutation and Poisson Competition schedulers with the bounds (4), (6) and (7) considered in this section. This is because the expression (5) is a limit and the expression (8) is for  $E_2^{ij}$ , not  $E_3^{ij}$ .

## 5 Conclusion

In this paper we have analyzed the latency of four decomposition-based schedulers for input-queued switches. We believe that there are a number of interesting open problems. First, it is possible that a tighter analysis of our framework of point processes could lead to better bounds for the existing schedulers, and may even motivate the construction of new ones. Second, we know of no non-trivial lower bounds on the best possible latency. It would be interesting to know what is the best value of  $\max_{ij} \rho_{ij} E_3^{ij}$  that can be achieved.

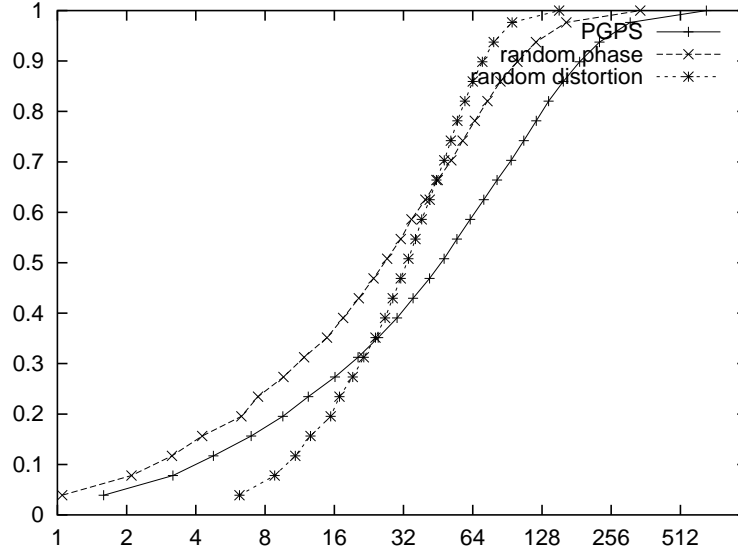


Figure 6: Fraction of  $ij$  pairs for which  $\rho_{ij} E_3^{ij} \leq x$ . The matrix has  $I = 64$ ,  $L = 4096$  and  $K = 2423$ .

## 6 Acknowledgements

The authors would like to thank Sem Borst, Jean-Yves Le Boudec and Lisa Zhang for many helpful discussions. The second author would like to thank the members of the Math Center at Bell Labs for their hospitality.

## References

- [1] C.S. Chang, W.J. Chen, and H.Y. Huang, "On service guarantees for input buffered crossbar switches: A capacity decomposition approach by Birkhoff and von Neumann," in *Proc. of IEEE IWQoS*, London, UK, 1999.
- [2] C.S. Chang, W.J. Chen, and H.Y. Huang, "Birkhoff-von Neumann input buffered crossbar switches," in *Proc. of IEEE INFOCOM'00*, Tel-Aviv, Israel, March 2000.
- [3] S. Li and N. Ansari, "Input-queued switching with QoS guarantees," in *Proc. of IEEE INFOCOM '99*, New York, NY, March 1999, pp. 1152 – 1159.
- [4] G. Birkhoff, "Tres observaciones sobre el algebra lineal," *Univ. Nac. Tucumán Rev. Ser. A5*, pp. 147 – 150, 1946.
- [5] J. von Neumann, "A certain zero-sum two-person game equivalent to the optimal assignment problem," *Contributions to the Theory of Games*, vol. 2, pp. 5 – 12, 1953.
- [6] B. Davie (Editor), A. Charny, F. Baker, J. Bennett, K. Benson, J.-Y. Le Boudec, A. Chiu, W. Courtney, S. Davari, V. Firoiu, C. Kalmanek, K. K. Ramakrishnan, and D. Stiliadis, "An expedited forwarding PHB," April 2001.
- [7] N. W. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proc. of IEEE INFOCOM*, 1996.

- [8] A. Mekikittikul and N. W. McKeown, "A practical algorithm to achieve 100% throughput in input-queued switches," in *Proc. of IEEE INFOCOM*, 1998.
- [9] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936 – 1948, December 1992.
- [10] J. Dai and B. Prabhakar, "The throughput of data switches with and without speedup," in *Proc. of IEEE INFOCOM '00*, Tel Aviv, Israel, March 2000, pp. 556 – 564.
- [11] E. Leonardi, M. Mellia, F. Neri, and M. Ajmone Marsan, "Bounds on average delays and queue size averages and variances in input-queued cell based switch," in *Proc. of IEEE Infocom 2001*, Anchorage, AK, 2001.
- [12] Devavrat Shah and Milind Kopikare, "Delay bounds for approximate maximum weight matching algorithms for input queued switches," in *Proc. of IEEE Infocom 2002*, New York, USA, June 2002.
- [13] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single node case," *IEEE/ACM Trans. on Networking*, vol. 1-3, pp. 344–357, June 1993.
- [14] J. Bennett, K. Benson, A. Charny, W. Courtney, and J.-Y. Le Boudec, "Delay jitter bounds and packet scale rate guarantee for expedited forwarding," *IEEE/ACM Trans. on Networking*, vol. 10, no. 4, pp. 529–540, August 2002.
- [15] J.-Y. Le Boudec and P. Thiran, *Network Calculus*, Springer-Verlag, 2001, (also available on-line at [http://icalwww.epfl.ch/PS\\_files/NetCal.htm](http://icalwww.epfl.ch/PS_files/NetCal.htm)).
- [16] R. Motwani and P. Raghavan, *Randomized algorithms*, Cambridge University Press, 1995.
- [17] Ward Whitt, *Stochastic-Processes Limits*, Springer, 2002.
- [18] H. Takagi, "Analysis and application of polling models," in *Performance Evaluation: Origins and Directions, Lecture Notes in Computer Science 1769*, G. Haring, C. Lindemann, and M. Reiser, Eds., 2000, pp. 423–442.
- [19] A. Hung, G. Kesidis, and N. McKeown, "ATM input-buffered switches with the guaranteed rate property," in *Proc. of IEEE ISCC*, 1998, pp. 331–335.
- [20] T. Lee and C. Lam, "Path switching – A quasi-static routing scheme for large scale ATM packet switches," *IEEE Journal on Selected Areas of Communications*, vol. 15, pp. 914 – 924, 1997.
- [21] S. Chuang, A. Goel, N. McKeown, and B. Prabhakar, "Matching output queueing with combined input and output queueing," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 6, pp. 1030–1039, 1999.
- [22] A. Charny, P. Krishna, N. Patel, and R. Simcoe, "Algorithms for providing bandwidth and delay guarantees in input-buffered crossbars with speedup," in *Proc. of IEEE IWQoS*, Napa, CA, 1998.
- [23] I. Stoica and H. Zhang, "Exact emulation of an output queueing switch by a combined input output queueing switch," in *Proc. of IEEE IWQoS*, Napa, CA, 1998.
- [24] V. Tabatabaee, L. Georgiadis, and L. Tassiulas, "QoS provisioning and tracking fluid policies in input queueing switches," *IEEE/ACM Trans. on Networking*, vol. 9, no. 5, October 2001.
- [25] J. Giles and B. Hajek, "Scheduling multirate periodic traffic in a packet switch," in *1997 Conference on Information Sciences and Systems at John Hopkins University*, 1997.

- [26] Francois Baccelli and Pierre Brémaud, *Elements of Queueing Theory*, vol. 26, Applications of Mathematics, Springer-Verlag, 1991.
- [27] Wassily Hoeffding, “Probability inequalities for sums of bounded random variables,” *American Statistical Association Journal*, pp. 13–30, March 1963.
- [28] Pierre Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, Springer, 1999.
- [29] Annie Gravey, Jean-Raymond Louvion, and Pierre Boyer, “On the Geo/D/1 and Geo/D/1/n queues,” *Performance Evaluation, North-Holland*, vol. 11, pp. 117–125, 1990.
- [30] J. W. Roberts (Editor), *COST 224: Performance evaluation and design of multiservice networks*, Commission of the European Communities, 1991.
- [31] Jaroslav Hájek, Zbyněk Šidák, and Pranab K. Sen, *Theory of Rank Tests*, Academic Press, 1999.
- [32] J. L. Doob, “Heuristic approach to the kolmogorov-smirnov theorems,” *Ann. Math. Statist.*, vol. 20, pp. 293–403, 1949.
- [33] Bruce Hajek, “A queue with periodic arrivals and constant service rate,” in *Probability, Statistics and Optimization – a Tribute to Peter Whittle (F.P. Kelly ed., John Wiley and Sons*, pp. 147–158, 1994.

## A Proof of Proposition 6

**Proof:** Given that the schedule is stationary we can take any point  $n$ , hence, without loss of generality, we fix  $n$  to 0. Note that (2) is equivalent to:

$$\left\{ \max_{1 \leq m \leq L} [N_{ij}^-[T_0, T_m) - (1 - \rho_{ij})m] \leq \rho_{ij}E_2^{ij} \right\} \Leftrightarrow \left\{ \max_{1 \leq m \leq L} [\rho_{ij}m - N_{ij}[T_0, T_m)] \leq \rho_{ij}E_2^{ij} \right\}.$$

We note that  $\max_{1 \leq m \leq L} [\rho_{ij}m - N_{ij}[T_0, T_m)]$  is exactly the scaled version of Kolmogorov-Smirnov statistics.<sup>2</sup> It is a classical result, as  $L \rightarrow \infty$ , we have the convergence in distribution:

$$\max_{1 \leq m \leq L} [N_{ij}^-[T_0, T_m) - (1 - \rho_{ij})m] \Rightarrow \sqrt{\rho_{ij}(1 - \rho_{ij})L} \sup_{0 \leq t \leq 1} [B_0(t) - t].$$

Where  $B_0$  is Brownian bridge; the Gaussian process with  $E[B_0(t)] = t$  and  $\text{Cov}[B_0(s)B_0(t)] = s(t - s)$ ,  $0 \leq s \leq t \leq 1$ .

It is known (Doob [32])

$$P\left(\sup_{0 \leq t \leq 1} [B_0(t) - t] > b\right) = e^{-2b^2}.$$

From the above convergence and equating the last limit distribution with  $\varepsilon$ , we obtain the stated result.  $\square$

## B Proof of Proposition 7

**Proof:** First, note that (3) is equivalent to

$$\left\{ \max_{n \geq 0, m > 0} [\rho_{ij}m - N_{ij}[T_n, T_{n+m})] \leq \rho_{ij}E_3^{ij} \right\}.$$

From periodicity of  $N_{ij}[T_0, T_n)$ ,  $n > 0$ , it follows

$$\max_{n \geq 0, m > 0} [\rho_{ij}m - N_{ij}[T_n, T_{n+m})] = \max_{1 \leq k \leq m \leq 2L} [\rho_{ij}(m - k) - N_{ij}[T_k, T_m)].$$

Now, similarly as in the proof of Proposition 6, we have the weak limit (see Hajek [33])

$$\max_{1 \leq k \leq m \leq 2L} [\rho_{ij}(m - k) - N_{ij}[T_k, T_m)] \Rightarrow \sqrt{\rho_{ij}(1 - \rho_{ij})LM},$$

where  $M = \sup_{0 \leq s \leq t \leq 2} [B_0(t) - B_0(s)]$ , the range of the Brownian bridge. It is known (see Hajek [33] and references therein)

$$P(M \leq c) = 1 - \sum_{l \geq 1} (4\ell^2 c^2 - 1) e^{-2\ell^2 c^2}.$$

---

<sup>2</sup>Let  $z$  be a vector of  $L$  elements with  $\ell_{ij}$  1's, and other elements equal to zero. The Kolmogorov-Smirnov statistics  $K_z^+$  is defined as follows (e.g. see [31], Section 6.3.7, and Whitt [17], Section 2.2),

$$K_z^+ := \frac{\max_{1 \leq m \leq L} [\rho_{ij}m - N_{ij}[T_0, T_m)]}{\sqrt{\sum_{l=1}^L (z_l - \rho_{ij})^2}}.$$



Suppose  $E_3^{ij}$  is given by (14) for some constant  $A$ . From the above weak limit we conclude

$$\mathbb{P}(\max_{1 \leq k \leq m \leq 2L} [\rho_{ij}(m-k) - N_{ij}[T_k, T_m]] \leq \rho_{ij} E_3^{ij}) \rightarrow 1 - \sum_{l \geq 1} (4\ell^2 A - 1) e^{-2\ell^2 A}.$$

Requiring the right-hand side is equal to  $1 - \varepsilon$ , for some fixed  $\varepsilon > 0$ , we recover the desired statement.  $\square$

## C Proof of Lemma 8

We now prove Lemma 8.

**Proof:** Suppose inductively that we have already chosen  $y_1, \dots, y_{v-1}$  such that,

$$\sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}] \leq \sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)].$$

This can trivially be done for  $v = 1$ . Then,

$$\begin{aligned} & \sum_y \sum_i \mathbb{P}(Y_v = y | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}) \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y] \\ &= \sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}] \leq \\ &\leq \sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)]. \end{aligned}$$

Since  $\sum_y \mathbb{P}(Y_v = y | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}) = 1$ , there exists a fixed value  $y_v$  such that,

$$\sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y_v] \leq \sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)].$$

Then, by Inequality (16),

$$\begin{aligned} \mathbb{P}(\sigma_i | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y_v) &\leq \sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y_v] \\ &\leq \sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j)]. \end{aligned}$$

The proof follows by induction.  $\square$

To find  $y_v$  we minimize,

$$\sum_i \mathbb{E}[\prod_{j=1}^{n_2} f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y] = \sum_i \prod_{j=1}^{n_2} \mathbb{E}[f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y],$$

over all possible values of  $y$ . We can exchange the expectation with the product due to the independence of the  $X_j$ . Hence we only need to be able to calculate the  $\mathbb{E}[f_{ij}(X_j) | Y_1 = y_1, \dots, Y_{v-1} = y_{v-1}, Y_v = y]$  in isolation. This is feasible in all our applications of Lemma 8.

## D Proof of Lemma 11

**Proof:** First note,

$$\begin{aligned}
P(N_{ij}[T_n, T_{n+m}) = \ell) &= P(N_{ij}\{T_n\} = 0, N_{ij}(T_n, T_{n+m}) = \ell) + \\
&\quad + P(N_{ij}\{T_n\} = 1, N_{ij}(T_n, T_{n+m}) = \ell - 1) \\
&= (1 - \rho_{ij})P(N_{ij}(T_n, T_{n+m}) = \ell) + \rho_{ij}P(N_{ij}(T_n, T_{n+m}) = \ell - 1).
\end{aligned}$$

We use an elementary result that a superposition of Poisson processes with finite intensities is Poisson process. In the last equality, we exploit the independence of increments of Poisson process.

Now, compute:

$$\begin{aligned}
P(N_{ij}(T_n, T_{n+m}) = \ell) &= \int_0^\infty P(N_{ij}(T_n, T_n + t) = \ell, N_{\bar{i}\bar{j}}(T_n, T_n + t) = m - 1) dt \\
&= \int_0^\infty P(N_{ij}(T_n, T_n + t) = \ell, N_{\bar{i}\bar{j}}(T_n, T_n + t) = m - 1 - \ell) dt \\
&= \int_0^\infty P(N_{ij}(0, t) = \ell, N_{\bar{i}\bar{j}}(0, t) = m - 1 - \ell) dt \\
&= \int_0^\infty P(N_{ij}(0, t) = \ell) P(N_{\bar{i}\bar{j}}(0, t) = m - 1 - \ell) dt \\
&= \frac{\rho_{ij}^\ell (1 - \rho_{ij})^{m-1-\ell}}{\ell! (m-1-\ell)!} \int_0^\infty t^{m-1} e^{-t} dt \\
&= \frac{(m-1)!}{(\ell-1)! (m-\ell)!} \rho_{ij}^\ell (1 - \rho_{ij})^{m-1-\ell},
\end{aligned}$$

where the second equality is from, for any  $\mathcal{I} \in \mathbb{R}$ ,  $N\mathcal{I} = N_{ij}\mathcal{I} + N_{\bar{i}\bar{j}}\mathcal{I}$ ; the third equality is by independence and stationarity of the increments of  $N_{ij}$  and  $N_{\bar{i}\bar{j}}$ ; the forth equality follows by the independence of  $N_{ij}$  and  $N_{\bar{i}\bar{j}}$ ; in the fifth equality we use  $P(N_{ij}(0, t) = \ell) = \frac{(\rho_{ij}t)^\ell}{\ell!} e^{-\rho_{ij}t}$ , by definition of Poisson process; and finally, in the last equality  $\int_0^\infty t^m e^{-t} dt = m!$ , for  $m$  an integer.

The statement of the lemma follows from the last two identities, by plugging the latter into the former.  $\square$